Алгоритм стохастического градиента и его применение в задачах машинного обучения

к.ф.-м.н. М. С. Ананьевский

(с.н.с. лаб. УСС, ИПМаш РАН)

02 мая 2024 г.

Математическая постановка задачи обучения нейронной сети для распознавания изображений

$\mathbb{X}^j \subset \mathbb{R}^m$, $j = 1, \dots, k$ – конечное количество классов изображений,
$\mathbb{X}^j \cap \mathbb{X}^s = \emptyset$, $j \neq s$ – изображение не может быть одновременно в двух классах

$f(w, x) : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ – функция нейронной сети, где
w – настраиваемые параметры

Требуется найти такое значение параметров $w = w_*$, чтобы

$$\forall j : \begin{cases} \forall x \in \mathbb{X}^j : f(w_*, x) = e_j, \\ \forall x \notin \mathbb{X}^j : f(w_*, x) \neq e_j, \end{cases} \tag{1}$$

где $e_j = (0, \dots, 0, 1, 0, \dots, 0)^T$ – единичный орт.

Математическая постановка задачи обучения нейронной сети для распознавания изображений

$f(w, x) : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ – функция нейронной сети, где
w – настраиваемые параметры

Для наперед заданного $\epsilon > 0$, требуется найти такое значение параметров $w = w_*$, чтобы

$$\forall j : \begin{cases} \forall x \in \mathbb{X}^j : ||f(w_*, x) - e_j|| < \epsilon, \\ \forall x \notin \mathbb{X}^j : ||f(w_*, x) - e_j|| > \epsilon. \end{cases} \tag{2}$$

(!) С практической т.з. у нас есть большой произвол в возможности модификации функции $f(w, x)$.

Математическая постановка задачи обучения нейронной сети для распознавания изображений

Задана обучающая выборка $\mathbb{X}_T = \mathbb{X}_T^1 \cup \dots \cup \mathbb{X}_T^k$.
$\mathbb{X}_T^j \subseteq \mathbb{X}^j \subset \mathbb{R}^m$, $j = 1, \dots, k$.

Требуется найти такое значение параметров $w = w_*$, чтобы

$$\forall j : \forall x \in \mathbb{X}_T^j : ||f(w_*, x) - e_j|| < \epsilon \tag{3}$$

(?) Как связано решение задачи (3) с решением задачи (2)?
(?) Каким свойствам должна удовлетворять обучающая выборка $\mathbb{X}_T$, чтобы решения сходились (по вероятности)?

(!) Выборка $\mathbb{X}^T$ может быть задана заранее, а может формироваться в процессе обучения (бесконечно).

Математическая постановка задачи обучения нейронной сети для распознавания изображений

Требуется решить задачу минимизации функционала

$$Q(w) = \sum_{j=1}^{k} \sum_{x \in \mathbb{X}_T^j} KL(f(w,x), \tilde{e}_j) \to \min_w \qquad (4)$$

здесь, $KL(\cdot,\cdot)$ – расстояние Кульбака–Лейблера, $\tilde{e}_j$ – чуть измененные орты (чтобы не было нулей).

(?) Как связано решение задачи (4) с решением задач (3), (2)?

Математическая постановка задачи обучения нейронной сети для распознавания изображений

Требуется решить задачу минимизации функционала

$$Q(w) = \sum_k \sum_i \sum_v g(w, x_{k,i,v}) \to \min_w \tag{5}$$

здесь, k соотвествует классу изображения, i – объекту изображения, v – ракурсу.

(!) Множество $\mathbb{X}_T$ обладает явно выраженными кластерами. С некоторыми оговорками можно сказать, что для почти всех w:

$$\forall v, v' : ||g(w, x_{k,i,v}) - g(w, x_{k,i,v'})|| << 1$$

Процедура Роббинса – Монро

Требуется найти значение $x \in \mathbb{R}$, при котором функция $g(x) = 0$.

Значение функции $g(\cdot)$ в точке x измеряется с центрированной помехой:

$$g(x) = E_\psi[G(\psi, x)] = \int_{\mathbb{R}} G(\psi, x) P_\psi(d\psi)$$

Алгоритм (x_0 – выбирается произвольно):

$$x_n = x_{n-1} - \alpha_n G(\hat{\psi}_n, x_{n-1})$$

Необходимо:

$$\sum_{n=1}^{+\infty} a_n = +\infty, \quad \sum_{n=1}^{+\infty} a_n^2 < +\infty$$

Процедура Роббинса – Монро и "батчи"

Требуется найти значение $x \in \mathbb{R}^n$, при котором функция $\nabla_x g(x) = 0$.

$$\nabla_x g(x) = E_\psi[\nabla_x G(\psi, x)] = \int_{\mathbb{R}} \nabla_x G(\psi, x) P_\psi(d\psi)$$

Алгоритм (x_0 – выбирается произвольно):

$$x_n = x_{n-1} - \alpha_n \nabla_x G(\hat{\psi}_n, x_{n-1})$$

$$\nabla_w Q(w) = E_\psi \left[\nabla_w \sum_{x \in \mathbb{X}_T^\psi} KL(f(w, x), \tilde{e}_j) \right], \tag{6}$$

где $\mathbb{X}_T^\psi$ – это случайно сформированное подмножество $\mathbb{X}_T$ ("батч").

Улучшение процедуры Роббинса – Монро для нашего функционала

(?) Какой оптимальный размер батча с т.з. скорости сходимости с учетом вычислительной стоимости подсчета градиента?

(?) Можно ли выбирать батч более оптимальным методом, например, минимизируя дисперсию градиента, чтобы улучшить сходимость? Использовать результаты предварительной кластеризации обучающей выборки?

(?) Имеет ли смысл предварительно решить задачу минимизации на множестве, сформированном из представителей кластеров обучающей выборки?

(?) Можно ли оптимизировать вычисления градиента, т.е. вычислять его не точно на батче, учитывая, что все равно это случайное приближение?